ICT CONFERENCE

Abstract book

Braga, 2023
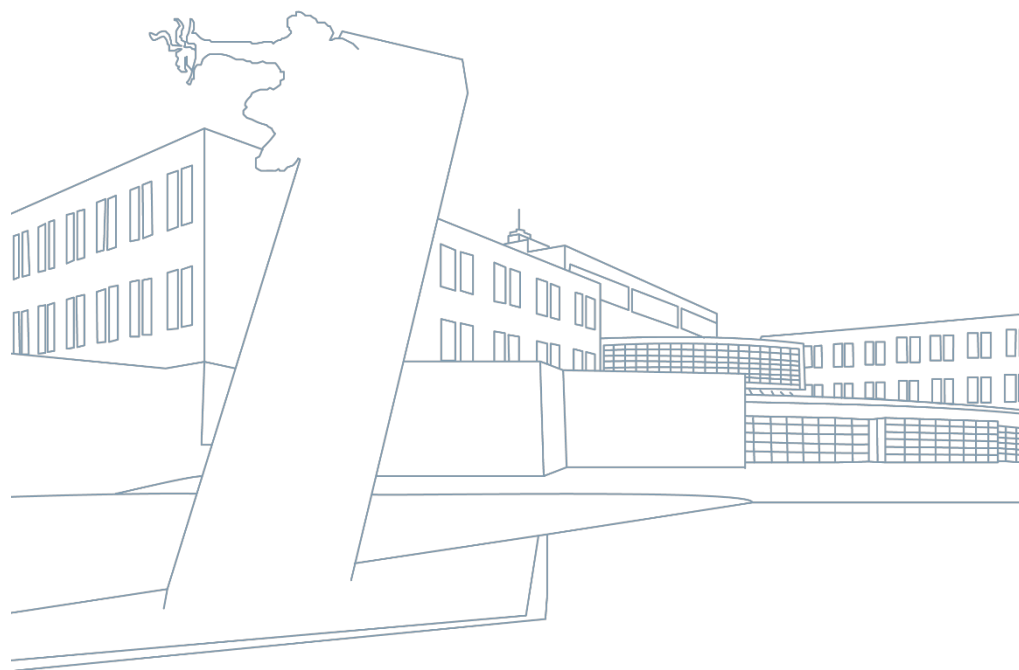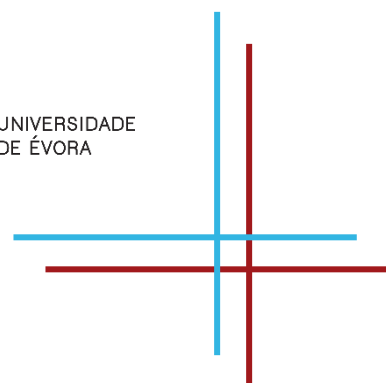
Statistical procedures applied to floods in the Douro River Basin

Carlos Ferreira[1], Cecília Castro[2], Jorge Pamplona[3]

[1] Universidade do Minho, Instituto de Ciências da Terra, Pólo do Minho, Departamento de Ciências da Terra, Escola de Ciências, Universidade do Minho, Campus de Gualtar , 4710-057 BRAGA, cmff.000@gmail.com

[2] Universidade do Minho, CMAT - Centro de Matemática, Departamento de Matemática, Escola de Ciências, Universidade do Minho, Campus de Gualtar , 4710-057 BRAGA, cecilia@math.uminho.pt

[3] Universidade do Minho, Instituto de Ciências da Terra, Pólo do Minho, Departamento de Ciências da Terra, Escola de Ciências, Universidade do Minho, Campus de Gualtar , 4710-057 BRAGA, cmff.000@gmail.com , jopamp@dct.uminho.pt

Abstract

The aim was to study flood event triggers. To this end, flood occurrence data were collected and stored, as well as hydro-meteorological variables for the Douro River basin.

The Douro River and its tributaries have very steep longitudinal profiles in some sections, and consequently sudden rises in water levels are observed after heavy precipitation.

The data treatment and analysis begins with a univariate study of the different variables. Several statistical procedures are used, in order to understand the possible relationship of each of the observed factors with the occurrence of floods, either individually or globally. This is done using Fisher's exact tests, chi-square tests, logistic regression models, and random forests explaining the flood phenomenon, adjusted on the basis of available data.

In the logistic regression model, there is a need to use the categorized predictors because their empirical distributions exhibit very sharp positive skewness, with many outliers. In this model, the important predictors are monthly-accumulated precipitation (mm) and monthly surface discharge (dam^3). The model has a specificity of over 90% but sensitivity of only 33.3%, which is not surprising given the complexity of the phenomenon under analysis. The discriminatory ability of the logistic regression model, measured by the area under the ROC curve, AUC, is 76.8% and is therefore acceptable.

The random forest algorithm is used with the uncategorized variables, since it does not depend on their distributions. With the same predictors, specificity higher than 99% and a sensitivity of only 60% is obtained with this procedure, indicating an excellent performance taking into account the complexity of the phenomenon and the fact that only two predictors are being used.

Keywords: Douro River basin, floods, statistical procedures